*Send Orders for Reprints to reprints@benthamscience.net*

4 *Current Pharmaceutical Design*, 2014, *20*, 4-10

# Differential Network Analysis in Human Cancer Research

Ryan Gill[a], Somnath Datta[b] and Susmita Datta[b,*]

[a]*Department of Mathematics, University of Louisville, Louisville, KY, 40292, USA;* [b]*Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, 40202, USA*

**Abstract:** A complex disease like cancer is hardly caused by one gene or one protein singly. It is usually caused by the perturbation of the network formed by several genes or proteins. In the last decade several research teams have attempted to construct interaction maps of genes and proteins either experimentally or reverse engineer interaction maps using computational techniques. These networks were usually created under a certain condition such as an environmental condition, a particular disease, or a specific tissue type. Lately, however, there has been greater emphasis on finding the differential structure of the existing network topology under a novel condition or disease status to elucidate the perturbation in a biological system. In this review/tutorial article we briefly mention some of the research done in this area; we mainly illustrate the computational/statistical methods developed by our team in recent years for differential network analysis using publicly available gene expression data collected from a well known cancer study. This data includes a group of patients with acute lymphoblastic leukemia and a group with acute myeloid leukemia. In particular, we describe the statistical tests to detect the change in the network topology based on connectivity scores which measure the association or interaction between pairs of genes. The tests under various scores are applied to this data set to perform a differential network analysis on gene expression for human leukemia. We believe that, in the future, differential network analysis will be a standard way to view the changes in gene expression and protein expression data globally and these types of tests could be useful in analyzing the complex differential signatures.

**Keywords:** Differential network analysis, human acute leukemia, permutation test, genetic modules.

## 1. INTRODUCTION

In the last decade there has been a large volume of literature on identifying or constructing gene and protein association or interaction networks for a specific biological condition either experimentally or computationally. Detecting an association or interaction network experimentally remains an expensive and labor-intensive process. The yeast two-hybrid system is the most commonly-used method to construct a network experimentally. It received a significant criticism for having a large number of false positives [1]. In spite of the criticism, using this experimental system, scientists have constructed association or interaction networks of many species such as Saccharomyces cerevisiae [2] and humans [3]. Mass spectrometry was also used to identify large scale experimental protein-protein interaction networks. Various authors [4-6] discussed the relative strengths and weaknesses of these methods and the consistency of the results using these different high-throughput techniques. Nevertheless, these experimental approaches remained popular for a while amongst scientists for constructing protein-protein and genetic association or interaction networks. However, due to the prohitive cost and labor intensive experimental methods, computational approaches to reverse engineer the protein and gene association networks became a lucrative alternative. In particular, our group came out with a novel computational method based on the partial least squares (PLS) regression technique for reconstruction of genetic networks from microarray data [7]. There had been other computational methods of reconstructing the association or interaction network as well [8-11].

However, all the above mentioned experimental and computational methods for constructing or reconstructing the gene and protein network were static in nature. In practice, however, biological systems are highly dynamic in response to perturbations caused by evolutionary changes, disease conditions, environmental stresses and result in changes the topology of the networks. Hence there is a growing interest and need for effective methods of examining the network structure under different biological settings and determining if there is a difference between the networks under two or more different experimental conditions or time points (differential networks analysis). In other words, differential network analysis inspects the networks under different systems to identify which parts of the network get affected by the perturbation to the system.

As mentioned above, there is a new development of studying the differential network experimentally and computationally under two or more experimental conditions. For example, [12] developed luminescence-based mammalian interactions mapping to identify protein-protein interaction (PPI) among a set of human factors under the presence and absence of stimulation by transforming growth factor beta. In this review however we are mostly concerned with computational approaches for measuring differential interaction networks. Weighted gene coexpression network analysis (WGCNA) of gene expression data was used in lean and heavy mice to identify the differences in connectivity and modular structure of the networks [13]. Unsupervised hierarchical clustering and supervised nonparametric procedures were used by [14] to examine differences in gene interactions and network structures for BCR-ABL positive and BCR-ABL negative adults with acute lymphoblastic leukemia (ALL). The response in the gene expression of lung epithelial cells to the H5N1 influenza virus was compared to response of those cells to the Respiratory Syncytial Virus and used to help cluster the genes using the Gene Ontology database with the human network of gene interactions and checked with a support vector machine (SVM) algorithm by [15]. Gene expressions of women with polycystic ovary syndrome (PCOS) with and without insulin resistance were compared with matched controls, and then genes that were determined to be differentially expressed were classified using ingenuity pathway analysis by [16]. Recently, [17] used epistatic miniarray profile in budding yeast to not only elucidate genetic interaction maps but also used differential interaction score for each gene pair under two different experimental conditions. In the remainder of this review/tuturial article we will concentrate on the computational methods proposed by our group [18] for conducting a differential network analysis.

*Address correspondence to this author at the Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, 40202, USA;
Tel: 5028520081; Fax: 5028523294; E-mail: susmita.datta@louisville.edu

1873-4286/14 $58.00+.00 

There are various ways to define/formulate what is meant by "differential network" or "difference between networks", though. Three definitions were considered in [18], and for each type of difference, a formal statistical test procedure was proposed based on a connectivity score which measures the strength of the association or interaction between selected pairs of genes. More specifically, these were (1) a test for differential connectivity of a single gene with other genes in two networks for two different biological systems, (2) a test for differential connectivity of a set of "interesting" genes, and (3) a test for differential modular structure between two networks.

The rest of this article is organized as follows. In the "computational methods" section, we describe several types of connectivity scores that can be used for network construction. We then review the three formal statistical tests proposed in [18] for differential analysis of networks under two experimental conditions. These tests are then applied to a well-known human acute leukemia data set of [19] (described in Section 3) based on the various connectivity scores to illustrate the computational pipeline. The illustrations and findings are described in Section 4 of the paper. The article concludes with a discussion section.

## 2. COMPUTATIONAL METHODS

This section describes a number of formal statistical methods proposed by [18] to test for a difference between two biological conditions. These tests include methods for assessing whether the connectivity of a particular gene differs between two networks, whether the connectivity of a set of genes differs between two networks, and whether the overall modular structures of the two networks are different. All of these tests are based on a set of connectivity scores which measure the strength of the association between pairs of genes in each of the networks.

The connectivity scores that are needed to construct the networks and to perform the statistical tests for a differential structure can be computed in many ways. We first describe four types of scores that are implemented in the contributed R package **dna** by [20] which is freely available from CRAN (http://CRAN.R-project.org/), and then we describe the statistical test procedures that use these scores.

### 2.1. Network Construction

Let $X$ be an $N\times p$ data matrix of gene expression values for $N$ subjects and $p$ genes. This subsection describes various commonly-used statistical methods that will be used for computing connectivity scores between each pair of genes in the network.

#### *2.1.1 Correlation*

The simplest measure of association between two variables is the correlation coefficient. Let $x_{ji}$ be the expression value for the $i$th gene of the $j$th subject and let $\bar{x}_i$ be the mean of the expression values among all subjects for the $i$th gene. Then the connectivity score for the $i$th and the $k$th genes based on the Pearsonian correlation coefficient is given by

$$s_{ik} = \frac{\sum_{j=1}^{N}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)}{\sqrt{\sum_{j=1}^{N}(x_{ji}-\bar{x}_i)^2 \sum_{j=1}^{N}(x_{jk}-\bar{x}_k)^2}}.$$

#### *2.1.2 Partial Least Squares*

Regression methods can be used to examine how the expression level for a specified gene is affected by each of the other genes simultaneously. Of course, the classical ordinary least squares (OLS) linear regression method is not applicable in this setting because the number of variables (genes) exceeds the number of subjects in such high throughput experiments. Dimension reduction (latent variable) methods such as partial least squares (PLS) can be used to obtain a smaller set of linear combinations of variables called PLS latent factors which can be used with an OLS linear regression model (Datta, 2001). The following PLS method for network construction described in Pihur et al. (2008) can be used to compute connectivity scores for the statistical test procedures discussed later.

Let $x_i$ be an $N$-dimensional vector of centered and scaled expression values for the $i$th gene. The linear model

$$x_i = \sum_{j=1}^{M}\beta_{ij}t_i^{(j)} + \text{error} \quad (1)$$

is used to model the expression values for the $i$th gene on $M$ PLS latent variables. The tuning parameter $M$ is selected by the user and the PLS latent variables are linear combinations of expressions of the remaining genes. The starting deflated design matrix $X_{-i}^{(1)}$ has columns $x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_p$, the $j$th deflated design matrix is determined using the iterative formula

$$X_{-i}^{(j)} = X_{-i}^{(j-1)} - t_i^{(j-1)}\left(t_i^{(j-1)}\right)^T X_{-i}^{(j-1)} / \left(\left(t_i^{(j-1)}\right)^T t_i^{(j-1)}\right)$$

for $j=2,\ldots,M$, the columns of $X_{-i}^{(j)}$ are denoted by $x_{-i,1}^{(j)},\ldots,x_{-i,i-1}^{(j)},x_{-i,i+1}^{(j)},\ldots,x_{-i,p}^{(j)}$, and the $j$th latent variable is

$$t_i^{(j)} = \sum_{\substack{k=1\\k\neq i}}^{M} c_{ik}^{(j)} x_{-i,k}^{(j)}, j=1,\ldots,M \quad (2)$$

where the coefficients

$$c_{ik}^{(j)} = \frac{\left(X_{-i}^{(j)}\right)^T x_i}{\sqrt{x_i^T X_{-i}^{(j)}\left(X_{-i}^{(j)}\right)^T x_i}}$$

give the contribution of gene $k$ to the $j$th latent variable $t_i^{(j)}$ after accounting for the relation explained by the earlier latent variables. The estimate of the regression coefficients for the effect of the $j$th latent variable on the $i$th gene is then computed using ordinary least squares estimates

$$\hat{\beta}_{ij} = \frac{\left(t_i^{(j)}\right)^T x_i}{\left(t_i^{(j)}\right)^T t_i^{(j)}}.$$

Combining (1) with (2), the coefficient for the effect of $x_k$ on $x_i$, accounting for the effect of the other genes, is $\sum_{j=1}^{M}\hat{\beta}_{ij}c_{ik}^{(j)}$.

Finally, the estimate of the PLS regression coefficients are symmetrized to obtain an overall connectivity score for genes $i$ and $k$

$$s_{ik} = \frac{\sum_{j=1}^{M}\hat{\beta}_{ij}c_{ik}^{(j)} + \sum_{j=1}^{M}\hat{\beta}_{kj}c_{ki}^{(j)}}{2}.$$

#### *2.1.3 Principal Components*

Principal components (PC) regression is another commonly-used variable reduction method based on derived inputs that can be used to obtain connectivity scores for network constructions and the associated statistical test procedures. As with PLS regression, we assume the columns of $X$ are centered and scaled. The latent factors for PC regression of $x_i$ on the remaining genes are obtained based on the eigenvalue decomposition of $X_{-i}$, the matrix $X$ with the $i$th column excluded. Specifically, the singular value decomposition of $X_{-i}$ has the form $X_{-i} = U_i D_i V_i^T$ where $U_i$ is an $N\times(p-1)$ orthogonal matrix, $V_i$ is a $(p-1)\times(p-1)$ orthogonal matrix whose columns are the eigenvectors of $X_{-i}$, and $D_i$ is a $(p-1)\times(p-1)$ diagonal matrix such that the $i$th diagonal element is the square of the $i$th largest eigenvalue of $D_i$. See Golub and Van Loan (1996) for more details on the singular value decomposition.

For this method, the $j$th latent variable is $z_i^{(j)} = X_{-i}V_i^{(j)}$ where $V_i^{(j)}$ is the $j$th column of$V_i$. The columns of $V_i$ are often referred to

as principal components of $X_{-i}$. Since the latent variables are orthogonal, the coefficient estimates for the regression model

$$x_i = \sum_{j=1}^{M} \theta_{ij} z_i^{(j)} + \text{error}$$

can be obtained by univariate regressions and be expressed as

$$\hat{\theta}_{ij} = \frac{\sum_{k=1}^{N} x_{ik} z_{jk}}{\sum_{k=1}^{N} z_{jk}^2}.$$

Then, the vector of coefficients for PC regression of $x_i$ on $x_1$, …, $x_{i-1}, x_{i+1}, \ldots, x_p$ is

$\hat{\beta}_{-i} = \sum_{j=1}^{M} \hat{\theta}_{ij} V_i^{(j)}$. Letting $\hat{\beta}_{1i}, \ldots, \hat{\beta}_{i-1,i}, \hat{\beta}_{i+1,i}, \ldots, \hat{\beta}_{pi}$ denote the elements of the column vector $\hat{\beta}_{-i}$, the average $s_{ik} = \frac{\hat{\beta}_{ik} + \hat{\beta}_{ki}}{2}$ can be used as an overall connectivity score for genes $i$ and $k$. See Hastie et al. (2009) for more details on PC regression.

#### *2.1.4 Ridge Regression*

Shrinkage methods are useful in the linear regression setting when the number of variables exceeds the number of observations. Ridge regression can be used to model $x_i$ as a function of the expression levels for the other genes by computing the estimates of the coefficient values $\beta_{1i}, \ldots, \beta_{i-1,i}, \beta_{i+1,i}, \ldots, \beta_{pi}$ which minimize the penalized sum of squares

$$\sum_{j=1}^{N} \left( x_{ji} - \sum_{\substack{k=1 \\ k \neq i}}^{p} \beta_{ki} x_{jk} \right)^2 + \sum_{\substack{k=1 \\ k \neq i}}^{p} \beta_{ki}^2 .$$

With $X_{-i}$ denoting the matrix $X$ with the $i$th column excluded, the estimates can be computed using the formula

$$\hat{\beta}_{-i} = \left( X_{-i}^T X_{-i} + \lambda I \right)^{-1} X_{-i}^T x_i$$

where $\hat{\beta}_{-i}$ is a $(p-1)$-dimensional vector with components $\hat{\beta}_{1i}, \ldots, \hat{\beta}_{i-1,i}, \hat{\beta}_{i+1,i}, \ldots, \hat{\beta}_{pi}$, $\lambda$ is a positive-valued complexity parameter, and $I$ is the identity matrix. To obtain an overall connectivity score which is symmetric, $s_{ik} = \frac{\hat{\beta}_{ik} + \hat{\beta}_{ki}}{2}$ is used. See Hastie et al. (2009) for more details on ridge regression.

### 2.2 Tests for Differential Networks

#### *2.2.1 Testing for a Difference in Connectivity for a Single Gene*

Often as an exploratory step, it is desirable to determine whether the connectivity of a single gene is significantly different in two networks. Such an analysis on a gene by gene basis may identify a set of genes that are impacted in the associated process leading to a change in the network structure.

Let $X_A$ and $X_B$ be $N_A \times p$ and $N_B \times p$ data matrices with expression values for networks A and B, respectively, and let $s_{ik}^A$ and $s_{ik}^B$ be the connectivity scores for genes $i$ and $k$ in networks A and B, respectively. Now, define a function $D$ which assigns a distance to measure the difference between scores $s^A$ and $s^B$. The choice of $D$ proposed in Gill et al. (2010) is the widely-used $L_1$-distance $D(s^A, s^B) = |s^A - s^B|$.

The test statistic suggested in Gill et al. (2010) for assessing whether gene $i$ is differentially expressed in networks A and B is the average distance between connectivity scores involving gene $i$,

$$d(i) = \frac{1}{p-1} \sum_{\substack{k=1 \\ k \neq i}}^{p} D(s_{ik}^A, s_{ik}^B).$$

To assess the significance of this statistic, the following permutation test can be used. Create an $(N_A + N_B) \times p$ matrix $\mathbb{X}$ of gene expression values with the $N_A$ rows of $X_A$ followed by the $N_B$ rows of $X_B$. Then, randomly permute the rows of $\mathbb{X}$ to create an $(N_A + N_B) \times p$ permuted matrix $\mathbb{X}^\pi$ where the first $N_A$ rows are denoted by $X_A^\pi$ and the last $N_B$ rows are denoted by $X_B^\pi$. Let $s_{ik}^{\pi,A}$ and $s_{ik}^{\pi,B}$ be the connectivity scores for genes $i$ and $k$ computed based on the matrices $X_A^\pi$ and $X_B^\pi$, respectively, and compute

$$d^\pi(i) = \frac{1}{p-1} \sum_{\substack{k=1 \\ k \neq i}}^{p} D(s_{ik}^{\pi,A}, s_{ik}^{\pi,B}).$$

This computation is repeated for $P$ random permutations, and an approximate p-value for testing the null hypothesis that the gene-gene interactions involving the $i$th gene are the same for both networks is the proportion of times

$$\frac{\sum I\{d^\pi(i) \geq d(i)\}}{P}$$

that the test statistic based on the permuted data is at least as large as the test statistic based on the observed data, where $I(A) = 1$ if $A$ is true and $I(A) = 0$ if $A$ is false. The null hypothesis is rejected if the p-value is sufficiently small.

#### *2.2.2 Testing for a Difference in Connectivity for a Set of Genes*

The connectivity scores can also be used to test whether an "interesting" set of genes in $\mathcal{F}$ differs between two networks. Let $f$ denote the number of genes in $\mathcal{F}$. The test statistic suggested in Gill et al. (2010) for assessing whether the set $\mathcal{F}$ is differentially expressed in networks A and B is the average connectivity of gene pairs in $\mathcal{F}$,

$$\Delta(\mathcal{F}) = \frac{1}{f(f-1)} \sum_{\substack{i,j \in \mathcal{F} \\ i \neq j}} D(s_{ij}^A, s_{ij}^B).$$

The permutation procedure is similar to the procedure described for the test of individual genes.

For each permuted matrix, compute

$$\Delta^\pi(\mathcal{F}) = \frac{1}{f(f-1)} \sum_{\substack{i,j \in \mathcal{F} \\ i \neq j}} D(s_{ij}^{\pi,A}, s_{ij}^{\pi,B})$$

and the approximate p-value is computed using

$$\frac{\sum I\{\Delta^\pi(\mathcal{F}) \geq \Delta(\mathcal{F})\}}{P}.$$

#### *2.2.3 Testing for a Difference in Connectivity in the Overall Modular Structure*

Often the modular structure is visualized with graphs that includes edges to connect gene-gene pairs which have strong interactions. The test described in this section is based on the following definition of a module proposed by Gill et al. (2010). A collection of genes $\mathcal{F}$ is referred to as a module if (1) there are at least $m$ elements in $\mathcal{F}$ and (2) given any two genes $f_1$ and $f_2$ in $\mathcal{F}$, there is a path of genes $f_1 = g_1, \ldots, g_k = f_2$ for some $k \geq 2$ such that $|s_{g_j g_{j+1}}| \geq \varepsilon$ for $j = 1, \ldots, k-1$ where $m$ is a minimum size parameter that controls how many genes are needed to be considered a module and $\varepsilon$ is a threshold connectivity parameter which controls how large the connectivity score between a pair of genes must be in magnitude in order to place an edge on the graph to connect the genes.

Using this definition of a module, the test statistic suggested in Gill et al. (2010) for assessing whether the modular structure differs between networks A and B is

$$\mathcal{N} = 1 - \sum_{g \in \mathcal{G}_0} \frac{\left| \mathcal{F}_{Aj(g)} \cap \mathcal{F}_{Bj(g)} \right|}{\left| \mathcal{F}_{Aj(g)} \cup \mathcal{F}_{Bj(g)} \right|} \quad (3)$$

where $\mathcal{F}_{Aj(g)}$ and $\mathcal{F}_{Bj(g)}$ are the modules in networks A and B, respectively, that contain gene $g$, $\mathcal{M}_A = \{\mathcal{F}_{A1}, \ldots, \mathcal{F}_{AJ_A}\}$ and $\mathcal{M}_B = \{\mathcal{F}_{A1}, \ldots, \mathcal{F}_{AJ_B}\}$ are the collections of all modules in networks A and B, respectively, and

$$\mathcal{G}_0 = \left(\bigcup_{j=1}^{J_A} \mathcal{F}_{Aj}\right) \cap \left(\bigcup_{j=1}^{J_B} \mathcal{F}_{Bj}\right)$$

is the set of genes that is in some module in both networks. If $\mathcal{G}_0$ is empty, the sum in (3) is taken to be 0. This statistic takes values between 0 (identical modular structure) and 1 (nothing in common) and measures the difference in modular structures in the two networks.

The permutation procedure is similar to the procedure described for the other tests. For each permuted matrix, compute

$$\mathcal{N}^{\pi} = 1 - \sum_{g \in \mathcal{G}_0^{\pi}} \frac{\left|\mathcal{F}_{Aj(g)}^{\pi} \cap \mathcal{F}_{Bj(g)}^{\pi}\right|}{\left|\mathcal{F}_{Aj(g)}^{\pi} \cup \mathcal{F}_{Bj(g)}^{\pi}\right|}$$

where $\mathcal{F}_{Aj(g)}^{\pi}$ and $\mathcal{F}_{Bj(g)}^{\pi}$ are the modules obtained from the permuted data, and $\mathcal{G}_0^{\pi}$ is the set of genes that is in some module in both networks for the permuted data. Then the approximate p-value is computed using

$$\frac{\sum I\{\mathcal{N}^{\pi} \geq \mathcal{N}\}}{P}$$

where the sum is taken over all $P$ permutations.

## 3. DATA

Human leukemia is a cancer affecting the bone marrow and blood; it is classified into four types based on rate of growth (acute or chronic) and cell type (lymphocytic or myeloid). Acute leukemia involves immature cells which multiply rapidly, while chronic leukemia involves partially mature but abnormal cells and progresses more slowly. Leukemia which begins in lymphocytes are referred to as lymphocytic; otherwise, a case which starts in another location such as red blood cells, platelets, or other types of white blood cells are referred to as myeloid leukemia [24]. Treatment types and survival rates vary greatly based on the type of leukemia; so it is important to correctly classify the type [25].

Thus, acute leukemia can be classified as either acute lymphocytic leukemia (ALL) or acute myeloid leukemia (AML). The original ALL/AML data set of [19] includes gene expression levels for 6817 genes measured by Affymetrix high-density oligonucleotide microarrays for bone marrow samples from two groups of leukemia patients (27 subjects with ALL and 11 subjects with AML). Three preprocessing steps were performed on the original data by the Bioconductor package **multtest** [26]. First, a threshold minimum of 100 and threshold maximum of 16,000 was used on all expression values. Next, genes were excluded if the ratio of the maximum and minimum expression values for the gene was less than 5 or if the difference between the maximum and minimum expression values is less than 500. Finally, a base 10 logarithmic transformation was applied to the data. The resulting dataset that is provided within the **multtest** package [26] and which is used in this paper includes 3051 genes.

This data set was originally analyzed by [19] discussing classification of acute leukemia cases as either ALL or AML. A classifier was designed based on the weighted average of the 50 genes most correlated with the responses. This 50-gene classifier was assessed using cross-validation on the training data and using an additional 34 independent samples (24 from bone marrow and 10 from peripheral blood samples) as test data.

The data was also used as an example for the program **rankgene** which implements various measures including the statistic from [19], the twoing rule, information gain, the gini index, max minority, the sum minority, the sum of variances, and a one-dimensional support vector machine to rank genes based on their ability to distinguish between classes [27]. The mathematical description of these measures is given at http://genomics10.bu.edu/yangsu/rankgene/.

Both the training and test sets were analyzed in one example in [28]. Various classifiers based on discrimination methods were used including linear discriminant analysis, maximum likelihood discrimination rules which also include diagonal quadratic discriminant analysis and diagonal linear discriminant analysis, nearest neighbor methods where the number of neighbors was chosen by cross-validation, a classification and regression tree with 10-fold cross-validation, and aggregate classifiers including bagging and boosting. It was found for this and other data sets that the simpler classifiers like the nearest neighbor method and the diagonal linear discriminant analysis worked very well in general.

The classification problem was also examined by [29] using the standard support vector machine (SVM) and a modified version (CSVM) using all of the genes to build the classifier. The SVM correctly classified all but two observations in the test data, and the CSVM correctly classified all observations.

Classification based on single genes and gene pairs classifiers were considered by [30] using the concept of canonical dependent degree. Leave-one-out-cross-validation was used to assess the performance of these simple models, and 13 gene pairs were identified which had high classification accuracy.

As stated before, in Section 4, we re-analyze the data in terms of change in the network structure between the two types of the disease following our own open source package **dna** [20].

## 4. RESULTS

All analyses in this section were performed within the R statistical software environment [31], and the tests were implemented using the contributed R package **dna** [20] based on the preprocessed *golub* dataset from the **multtest** package freely available from Bioconductor [32]. The computationally-intense calculations presented in this section were performed using the Cardinal Research Cluster at the University of Louisville.

We started with 3051 genes, but used univariate logistic regressions with type of leukemia as the response and individual gene as the explanatory variable, and we selected the 300 most significant genes with the smallest p-values and 300 least significant genes with the largest p-values. Using these 600 genes and the 38 subjects, tests for differences in network connectivity of each gene were performed based on the four types of connectivity scores described in Section 2. At a significance level of .01, 28 genes were significantly different in terms of their network connectivities between the two cancer subtypes based on correlation scores, 146 genes were significant based on PLS scores, 163 genes were significant based on PC scores, and 206 genes were significant based on RR scores. All tests used the $L_1$-distance and rescaled scores; all other default options in the **dna** package were used for each connectivity score. Table **1** lists the gene names and corresponding p-values for each type of connectivity score for the 19 genes that were significant at a level .01 based on all four types of scores.

Some of these genes have been connected with leukemia in the literature. The catenin, delta 1 gene (*CTNND1*) encodes a protein which has roles in signal transduction and adhesion between cells (Entrez Gene) and consequently could be associated with the aggressive phenotype of BCR-ABL-positive ALL [14]. The poly (ADP-ribose) polymerase 1 (*PARP1*) gene plays a role in many cellular processes including differentiation, proliferation, tumor transformation, and DNA damage recovery (Entrez Gene), and there is evidence that it is involved in a mechanism that may be important for the development of ALL [33]. The *IGFBP5* promoter is activated by the meningioma 1 (*MN1*) gene [34] which has been

**Table 1. List of genes that are significant based on tests for differential connectivity of an individual gene at level .01 using correlation, partial least squares regression, principal components regression, and ridge regression to compute connectivity scores. The corresponding p-values for these tests are also listed.**

| Gene | Cor | PLS | PC | RR | Gene | Cor | PLS | PC | RR |
|---|---|---|---|---|---|---|---|---|---|
| *CTNND1* | .000 | .001 | .000 | .000 | *RHOH* | .002 | .005 | .000 | .000 |
| *HG2705-HT2801_s_at* | .001 | .001 | .000 | .000 | *PARP1* | .001 | .005 | .000 | .002 |
| | | | | | *RCOR1* | .000 | .003 | .000 | .005 |
| *IGFBP5* | .001 | .001 | .000 | .001 | *HG3925-HT4195_s_at* | .001 | .004 | .000 | .005 |
| *CRM1* | .002 | .000 | .000 | .000 | | | | | |
| *SERPING1* | .002 | .000 | .000 | .000 | *FURIN* | .004 | .006 | .000 | .004 |
| *CTSL1* | .000 | .002 | .000 | .002 | *GLRX* | .007 | .000 | .001 | .000 |
| *RNH1* | .003 | .000 | .000 | .000 | *ANXA1* | .005 | .007 | .000 | .003 |
| *CTDSP2* | .003 | .001 | .001 | .000 | *NO55* | .009 | .000 | .000 | .000 |
| *FEZ2* | .003 | .001 | .000 | .001 | *LYZ* | .003 | .009 | .001 | .002 |
| *PPP1CC* | .005 | .000 | .002 | .000 | | | | | |

**Table 2. Values of test statistics and p-values for tests of differential connectivity of functional clusters using correlation, partial least squares regression, principal components regression, and ridge regression to compute connectivity scores.**

| Functional Cluster | Genes | Cor | PLS | PC | RR |
|---|---|---|---|---|---|
| Enzyme inhibitor activity | *ANXA1*, *FURIN*, *RHOH*, *RNH1*, *SERPING1* | .024 | .001 | .016 | .000 |
| Immune system development | *BAK1, CD79A, CD8A, FAS, SP3, CSF1, CSF3, FLT3LG, HOXA9, IFI16, IL7R, MYH9, NFKB2, PLEK, PSEN1, RHOH, PRKDC, TP53, RELB* | .030 | .028 | .000 | .013 |
| Regulation of programmed cell death | *BAK1, BCL2A1, BNIP2, FAS, MNT, NAE1, RB1CC1, ARHGAP4, ARHGEF2, SON, TRAF1, B4GALT1, ACTN1, AARS, ALDH1A3, ANXA1, AZU1, ABL1, CAT, CDK5, CDKN1A, CSTB, DAP, DAPK1, DYNLL1, FNTA, FOXO1, FURIN, HSPA1B, HMOX1, HMGB1, HERPUD1, INPP5D, IFI16, MAP3K11, MCL1, MPO, MYO18A, NOS3, NFKB1, NFKBIA, NR4A1, PPT1, PPIF, PRDX1, PIM1, PSEN1, PRNP, RAC1, SCRIB, SQSTM1, STAT1, BCLAF1, PRKDC, SNCA, TNFRSF25, TP53, YWHAZ, ATK1, ETS1* | .007 | .027 | .001 | .004 |
| SH2 domain | *JAK1, SHB, ABL1, CHN2, FES, INPPL1, INPP5D, LCP2, PTPN11, PTPN6, STAT1, STAT3* | .090 | .178 | .220 | .153 |
| Lymphocyte/mononuclear cell/leukocyte proliferation | *CD79A, SHB, CXCR4, CCND3, IL23A, IL7R, PRKCD, TP53* | .215 | .088 | .020 | .059 |

demonstrated to be important in predicting outcomes in patients with AML [35-39]. It should be noted that although *MN1* is not among the list of genes in Table **1**, it was significant at level .05 based on all methods of computing the connectivity scores (p-values are .002, .022, and .000 for PLS, PC, and ridge regression, respectively) except the correlation (p-value of .064). *FURIN* encodes a calcium-dependent serine endoprotease that efficiently cleaves precursor proteins at their paired basic amino acid processing sites, and this gene possibly plays a role in tumor progression [40] and treatment [41]. Lysozyme is encoded by the gene *LYZ* and was present at significantly elevated levels in AML patients [42]. The gene *CRM1* is involved in the active transport of tumor suppressors but functions differently in cancer when overexpressed with overactive transport, and a method of blocking this gene's export of tumor suppressor proteins has been proposed [43]. A meta-analysis of genome-wide association studies identified that the REST co-repressor 1 (*RCOR1*) gene has a statistically significant association with mean platelet volume and platelet count [44]. Also, a few of the genes, *RHOH* and *ANXA1*, have been connected with chronic forms of leukemia [45, 46].

Tests for differences in connectivity were also performed for the functional clusters of genes shown in Table **2**. First, DAVID [47] functional clustering was used post-hoc with the genes in Table **1**, and five genes *ANXA1*, *FURIN*, *RHOH*, *RNH1*, and *SERPING1* involved in enzyme inhibitor activity were identified as the highest rated cluster. The test for differential connectivity for a set of genes that was described in Section 2 was performed based on each of the four types of connectivity scores and the difference between the ALL and AML classes were significant at level .05 for

each of the scores. The p-values for these tests are shown in Table **2**.

We also applied DAVID functional clustering to the set of 600 genes and performed the test for differences in connectivity on several classes of genes that were identified with high enrichment scores. The connectivity scores for the function cluster of 19 genes labeled immune system development were significantly different for the ALL and AML classes based on all four types of connectivity scores at level .05 as shown in Table **2**. The cluster of 60 genes involved in regulation of programmed cell death shown in Table **2** were also significant at level .05 based on all four types of connectivity scores. Several genes (*TP53*, *IFI16*, *BAK1*, and *FAS*) were present in both the immune system development and the regulation of programmed cell death clusters. The gene *TP53* encodes the protein *p53* which functions as a tumor suppressor [40], and it is known that AML patients often have a *TP53* gene mutation which also is associated with a poor prognosis [48]. The *IFI16* gene is upregulated by *p53* [49] and is involved in cellular senescence [50]. Expression levels of *IFI16* have been associated with myeloid differentiation cells [51]. The *BCL2-antagonist/killer 1* (*BAK1*) gene encodes a protein which induces apoptosis [40]; exposure to cell stress leads to an interaction between this protein and *p53* [52]. The *FAS* gene is also involved in the process of apoptosis and can be activated by *p53* as described by [53]. The results for a couple of other clusters are shown in the last two rows of Table **2**, but the differences are not statistically significant based on most of the types of scores.

Tests for differences in overall modular structure were also performed for the 600 genes and 38 subjects based on the four types of connectivity scores described in Section 2. All tests used minimum size parameter $m = 5$ and threshold connectivity parameter $\varepsilon = .6$ with the $L_1$-distance and rescaled scores; all other default options in the **dna** package were used for each connectivity score. The tests based on all four connectivity scores fail to reject the null hypothesis and the observed value of the test statistic for a difference in overall modular structure is not statistically significant. It should be noted however that this simply indicates that the changes in the network are more subtle and are not detected by changes in large highly connected modules in the two networks.

## 5. CONCLUSION/DISCUSSION

In this article we illustrated the concept of differential network analysis using high-throughput gene expression data. Although we briefly reviewed others' work in this topic, we mainly concentrate on our recent work on differential network analysis [18] in this tutorial. We used a well known and well studied human leukemia data set [19] to demonstrate our methods and to check the biological relevance of our computational methods as they apply to cancer research.

It is evident that although statistical tests for detecting significant differential connectivity of individual genes in two different leukemia samples using different association scores such as correlation scores or PLS scores obtain different results, the genes which were commonly differentially connected in two networks in all the methods are highly biologically relevant. We have also shown that differential connectivity of the clusters of genes in the first three important cancer related functional clusters are significantly different in two different types of leukemia samples irrespective of different computational methods. The approach based on the overall modular structure was not so successful for this data set. Overall we demonstrate that although network construction in a static manner individually for every type of disease or biological system is important, differential network analysis provides a more comprehensive view of the complex dynamic nature of a complex disease like cancer.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

## ACKNOWLEDGEMENTS

This research work was partially supported by NIH-CA133844 (Su. Datta).

## REFERENCES

[1] Futschik ME, Chaurasia G, Herzel H. Comparison of human protein-protein interaction maps. Bioinformatics 2007; 23: 605-11.

[2] Uetz P, Giot L, Cagney G, *et al*. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000; 403: 623-7.

[3] Rual JF, Venkatesan K, Hao T, *et al*. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005; 437: 1173-8.

[4] Ho Y, Gruhler A, Heilbut A, *et al*. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 2002; 415: 180-3.

[5] Bader JS. Greedily building protein networks with confidence. Bioinformatics 2003; 19: 1869-74.

[6] Berggard T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. Proteomics 2007; 7: 2833-42.

[7] Pihur V, Datta S, Datta S. Reconstruction of genetic association networks from microarray data: A partial least squares approach. Bioinformatics 2008; 24: 561-8.

[8] Basso K, Margolin AA, Stolovitzky G, *et al*. Reverse engineering of regulatory networks in human B cells. Nat Genet 2005; 37: 382-90.

[9] Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 2005; 21: 754-64.

[10] Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 2005; 4: 32.

[11] Yu J, Smith VA, Wang PP, *et al*. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 2004; 20: 3594-603.

[12] Barrios-Rodiles M, Brown KR, Ozdamar B, *et al*. High-throughput mapping of a dynamic signaling network in mammalian cells. Science 2005; 307: 1621-5.

[13] Fuller TF, Ghazalpour A, Aten JE, *et al*. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome 2007; 18: 463-72.

[14] Juric D, Lacayo NJ, Ramsey MC, *et al*. Differential gene expression patterns and interaction networks in BCR-ABL-positive and -negative adult acute lymphoblastic leukemias. J Clin Oncol 2007 25: 1341-9.

[15] Tatebe K, Zeytun A, Ribeiro RM, *et al*. Response network analysis of differential gene expression in human epithelial lung cells during avian influenza infections. BMC Bioinformatics 2010; 11: 170.

[16] Kaur S, Archer KJ, Devi MG, *et al*. Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis. J Clin Endocrinol Metab 2012; 97: E2016-21.

[17] Bandyopadhyay S, Mehta M, Kuo D, *et al*. Rewiring of genetic networks in response to DNA damage. Science 2010; 330: 1385-9.

[18] Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics 2010; 11: 95.

[19] Golub TR, Slonim DK, Tamayo P, *et al*. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999; 286: 531-7.

[20] Gill R, Datta S, Datta S. dna: Differential Network Analysis. R package version 1.0-0.

[21] Datta S. Exploring relationships in gene expressions: a partial least squares approach. Gene Expr 2001; 9: 249-55.

[22] Golub GH, Van Loan CF. Matrix computations. 3rd ed. Baltimore: Johns Hopkins University Press 1996.

[23] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction. 2nd ed. New York: Springer 2009.

[24] Leukemia -- Acute Lymphocytic [homepage on the internet]. American Cancer Society, Inc.; [updated 2013 Jan 18; cited 2013 Jan 26]. Available from http://www.cancer.org/cancer/leukemia-acutelymphocyticallinadults/detailedguide/leukemia-acute-lymphocytic-what-is-all.

[25] American Cancer Society. Cancer Facts and Figures 2012. American Cancer Society, Inc. 2012.

[26] Pollard KS, Gilbert HN, Ge Y, *et al*. multtest: Resampling-based multiple hypothesis testing, R package version 2.14.0.

[27] Su Y, Murali TM, Pavlovic V, *et al*. RankGene: identification of diagnostic genes based on expression data. Bioinformatics 2003; 19: 1578-9.

[28] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002; 97: 77-87.

[29] Zhang X, Ke H. ALL/AML cancer classification by gene expression data using SVM and CSVM approach. Genome Inform 2000; 11: 237-9.

[30] Wang X, Gotoh O. Accurate molecular classification of cancer using simple rules. BMC Med Genomics 2009; 2: 64.

[31] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

[32] Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 2004; 5: R80.

[33] Kannan S, Fang W, Song G, *et al*. Notch/HES1-mediated PARP1 activation: a cell type-specific mechanism for tumor suppression. Blood 2011; 117: 2891-900.

[34] Meester-Smoor MA, Moljin AC, Zhao Y, *et al*. The MNI ontoprotein activates transcription of the IGFBP5 promoter through a CACCC-rich consensus sequence. J Mol Endocrinol 2007; 38: 113-25.

[35] Langer C, Marcucci G, Holland KB, *et al*. Prognostic importance of MN1 transcript levels, and biologic insights from MN1-associated gene and microRNA expression signatures in cytogenetically normal acute myeloid leukemia: a cancer and leukemia group B study. J Clin Oncol 2009; 27: 3198-204.

[36] Schroeder T, Czibere A, Zohren F, *et al*. Meningioma 1 gene is differentially expressed in CD34 positive cells from bone marrow of patients with myelodysplastic syndromes with the highest expression in refractory anemia with excess of blasts and secondary acute myeloid leukemia. Leuk Lymphoma 2009; 50: 1043-6.

[37] Grosveld GC. MN1, a novel player in human AML. Blood Cells Mol Dis 2007; 39: 336-9.

[38] Pardee TS. Overexpression of MN1 confers resistance to chemotherapy, accelerates leukemia onset, and suppresses p53 and Bim induction. PLoS One 2012; 7: e43185.

[39] Heuser M, Beutel G, Krauter J, *et al*. High meningioma 1 (MN1) expression as a predictor for poor outcome in acute myeloid leukemia with normal cytogenetics. Blood 2006; 108: 3898-905.

[40] Entrez Gene. [homepage on the internet]. National Center for Biotechnology Information; [cited 2013 Jan 30]. Available from www.ncbi.nlm.nih.gov/gene?db=gene.

[41] Abi-Habib RJ, Liu S, Bugge TH, *et al*. A urokinase-activated recombinant diphtheria toxin targeting the granulocyte-macrophage colony-stimulating factor receptor is selectively cytotoxic to human acute myeloid leukemia blasts. Blood 2004; 104: 2143-8.

[42] Alsabti E. The prognostic value of serum lysozyme activity in acute myelogenous leukemia. Med Pediatr Oncol 1979; 6: 189-93.

[43] Ranganathan P, Yu X, Na C, *et al*. Preclinical activity of a novel CRM1 inhibitor in acute myeloid leukemia. Blood 2012; 120: 1765-73.

[44] Gieger C, Radhakrishnan A, Cvejic A, *et al*. New gene functions in megakaryopoiesis and platelet formation. Nature 2011; 480: 201-8.

[45] Troeger A, Johnson AJ, Wood J, *et al*. RhoH is critical for cell-microenvironment interactions in chronic lymphocytic leukemia in mice and humans. Blood 2012; 119: 4708-18.

[46] Falini B, Tiacci E, Liso A, *et al*. Simple diagnostic assay for hairy cell leukaemia by immunocytochemical detection of annexin A1 (ANXA1). Lancet 2004; 363: 1869-70.

[47] Dennis G, Jr., Sherman BT, Hosack DA, *et al*. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003; 4: P3.

[48] Bowen D, Groves MJ, Burnett AK, *et al*. TP53 gene mutation is frequent in patients with acute myeloid leukemia and complex karyotype, and is associated with very poor prognosis. Leukemia 2009; 23: 203-6.

[49] Song LL, Alimirah F, Panchanathan R, *et al*. Expression of an IFN-inducible cellular senescence gene, IFI16, is up-regulated by p53. Mol Cancer Res 2008; 6: 1732-41.

[50] Xin H, Pereira-Smith OM, Choubey D. Role of IFI 16 in cellular senescence of human fibroblasts. Oncogene 2004; 23: 6209-17.

[51] Dawson MJ, Trapani JA. IFI 16 gene encodes a nuclear protein whose expression is induced by interferons in human myeloid leukaemia cell lines. J Cell Biochem 1995; 57: 39-51.

[52] Leu JI, Dumont P, Hafey M, *et al*. Mitochondrial p53 activates Bak and causes disruption of a Bak-Mcl1 complex. Nat Cell Biol 2004; 6: 443-50.

[53] Haupt S, Berger M, Goldberg Z, *et al*. Apoptosis - the p53 network. J Cell Sci 2003; 116: 4077-85.

**Received: March 6, 2013** **Accepted: March 18, 2013**